# Capstone Project-2

# Yes Bank Stock Closing Price Prediction

## Team Members

Mohd Danish
Huzaifa Khan
Arbaaz Malik
Abdul Rahman Talha

## *Let's Predict!*

1. Overview & Objective
2. Data Pipeline
3. EDA
4. Regression Analysis
    a) Linear Regression
    b) Ridge Regression
    c) Lasso Regression
    d) ElasticNet Regression
5. Conclusion

# Overview & Objective

## Overview

Yes Bank is a well-known bank in the Indian financial domain. Since 2018, it has been in the news because of the fraud case involving Rana Kapoor. Owing to this fact, it was interesting to see how that impacted the stock prices of the company and whether Time series models or any other predictive models can do justice to such situations..

## Objective

This dataset has monthly stock prices of the bank since its inception and includes closing, starting, highest, and lowest stock prices of every month. The main objective is to predict the stock's closing price of the month.

# Data Pipeline

- **Data Preprocessing:** At this stage, we check for duplicate values and missing values and treat them if any. Furthermore, we check the datatype of the features present in our dataset and transform them if necessary
- **Exploratory Data Analysis (EDA):** At this stage, we conduct an EDA on the selected features in order to better understand their spread, pattern and relationship with the other features. It gives us an intuition as to what is going on in the dataset.
- **Model Building:** At this stage, we apply various models to understand which one will give us the best result.

# Data Summary

We have Yes Bank monthly stock price dataset. It has following features (Columns):

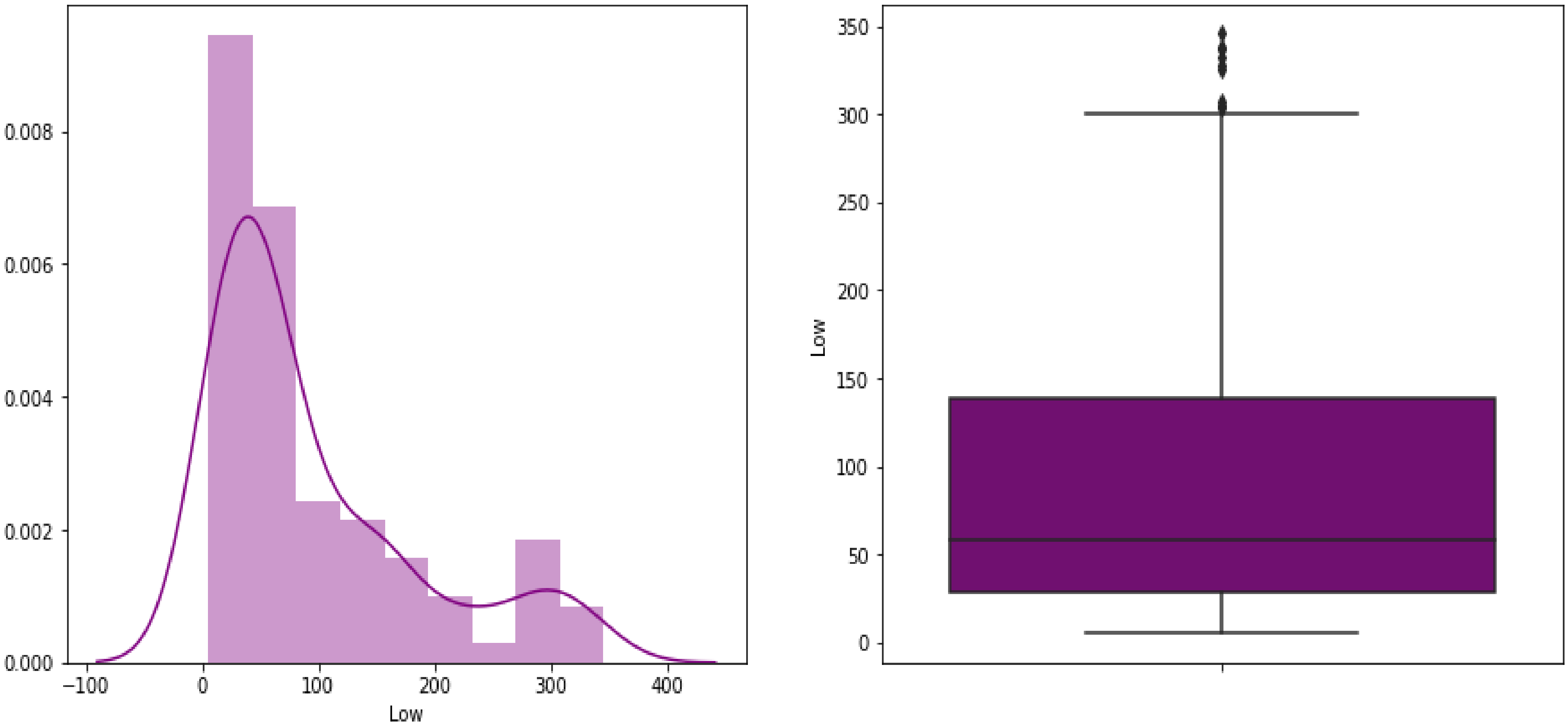
1) **Open** : Opening price of the stock of particular day
2) **High** : It's the highest price at which a stock traded during a period
3) **Low** : It’s the lowest price at which stock traded during a period
4) **Close** : Closing price of a stock at the end of a Trading Day
5) **Date** : We will use it as a index

Note: ‘Close’ will be our Dependent variable & Others will be independent.

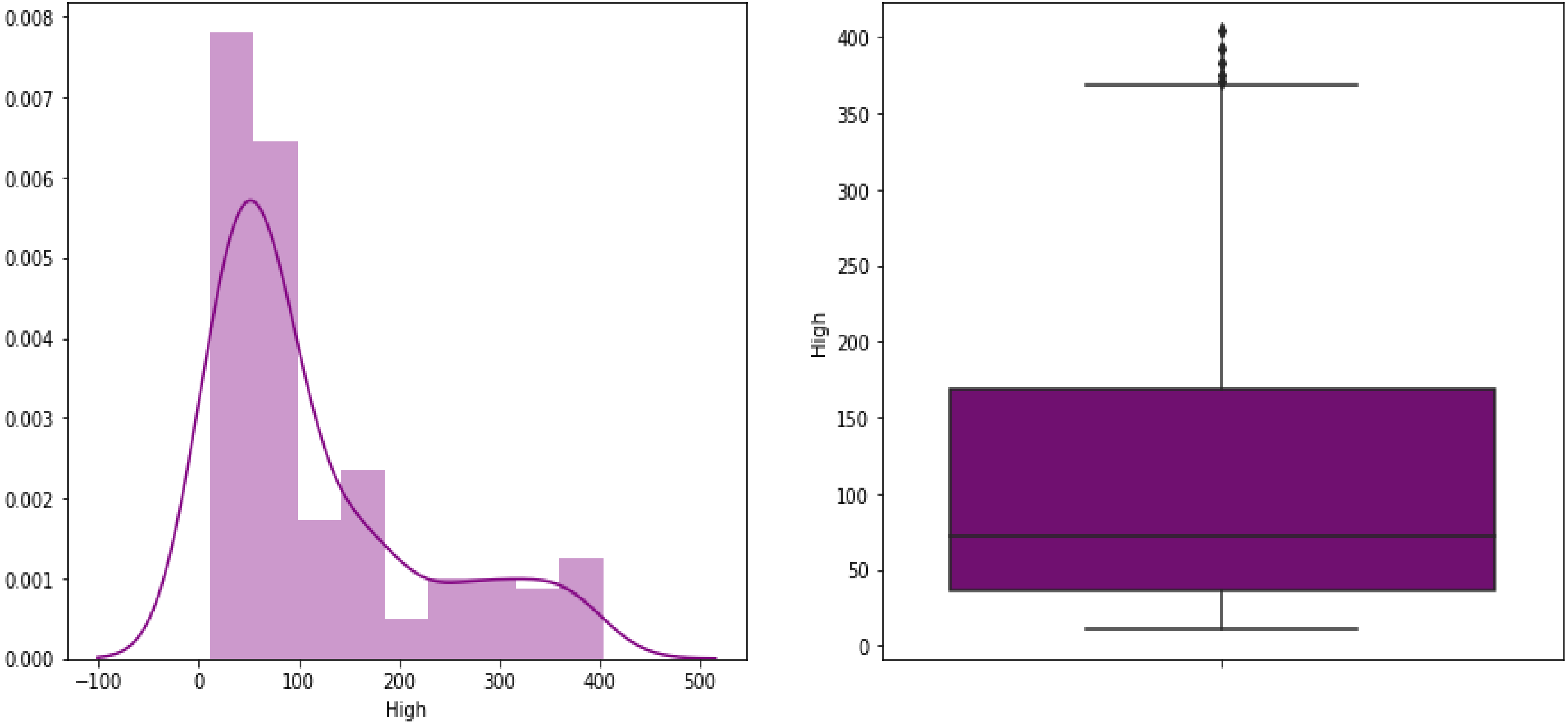
# Exploratory Data Analysis(EDA)

## Let’s Visualize dependent variable

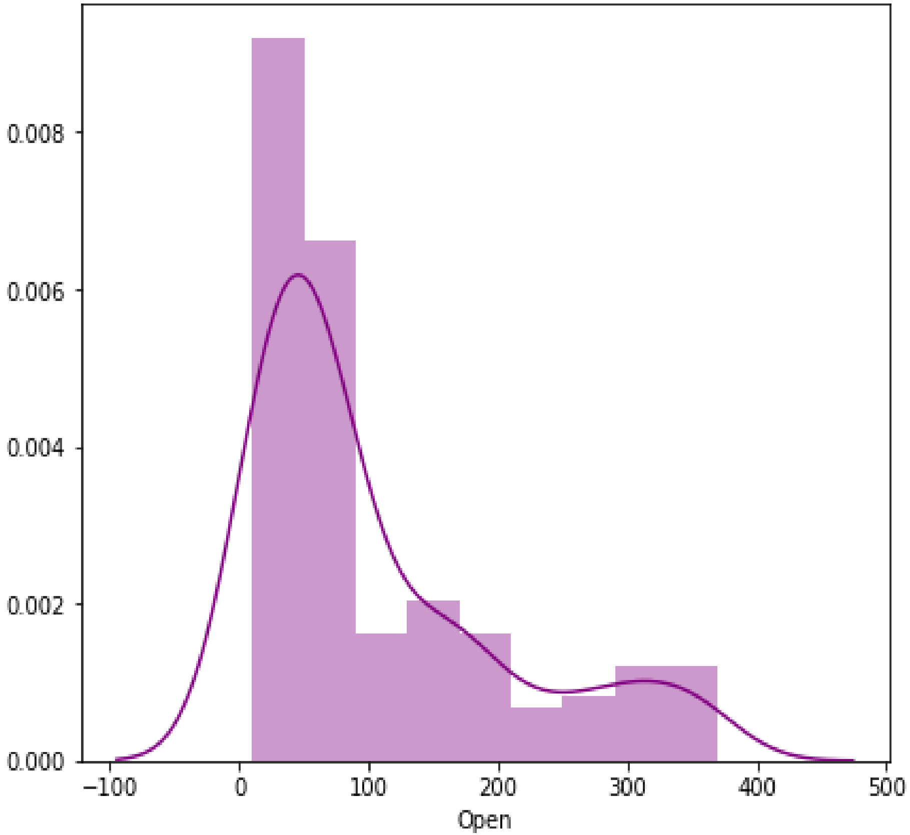
# Distribution of ‘Low’

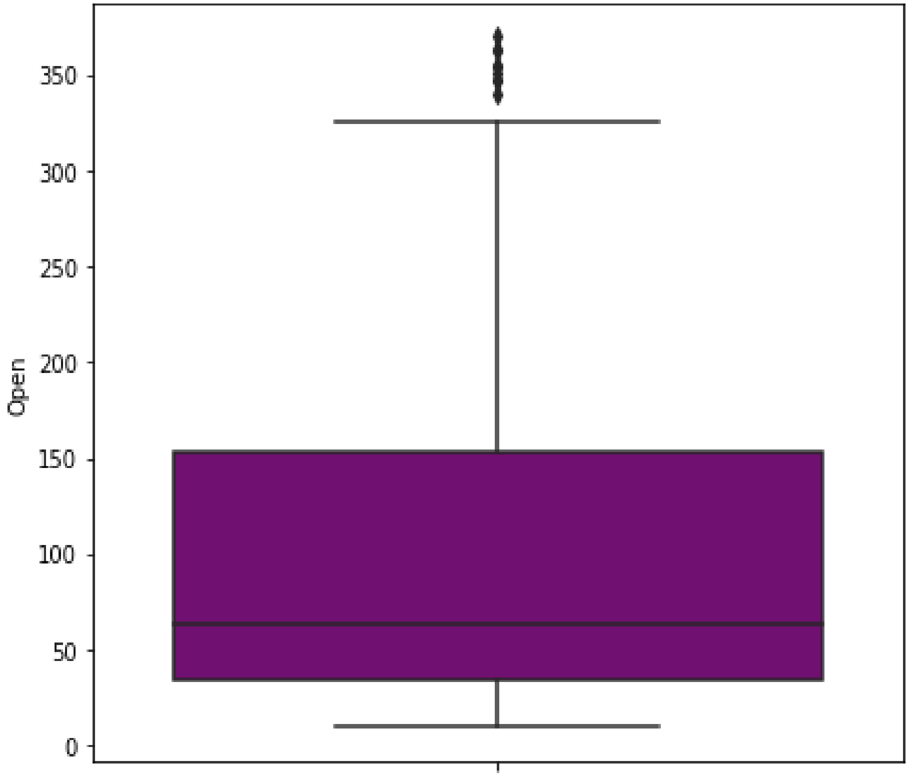


# Distribution of ‘High’

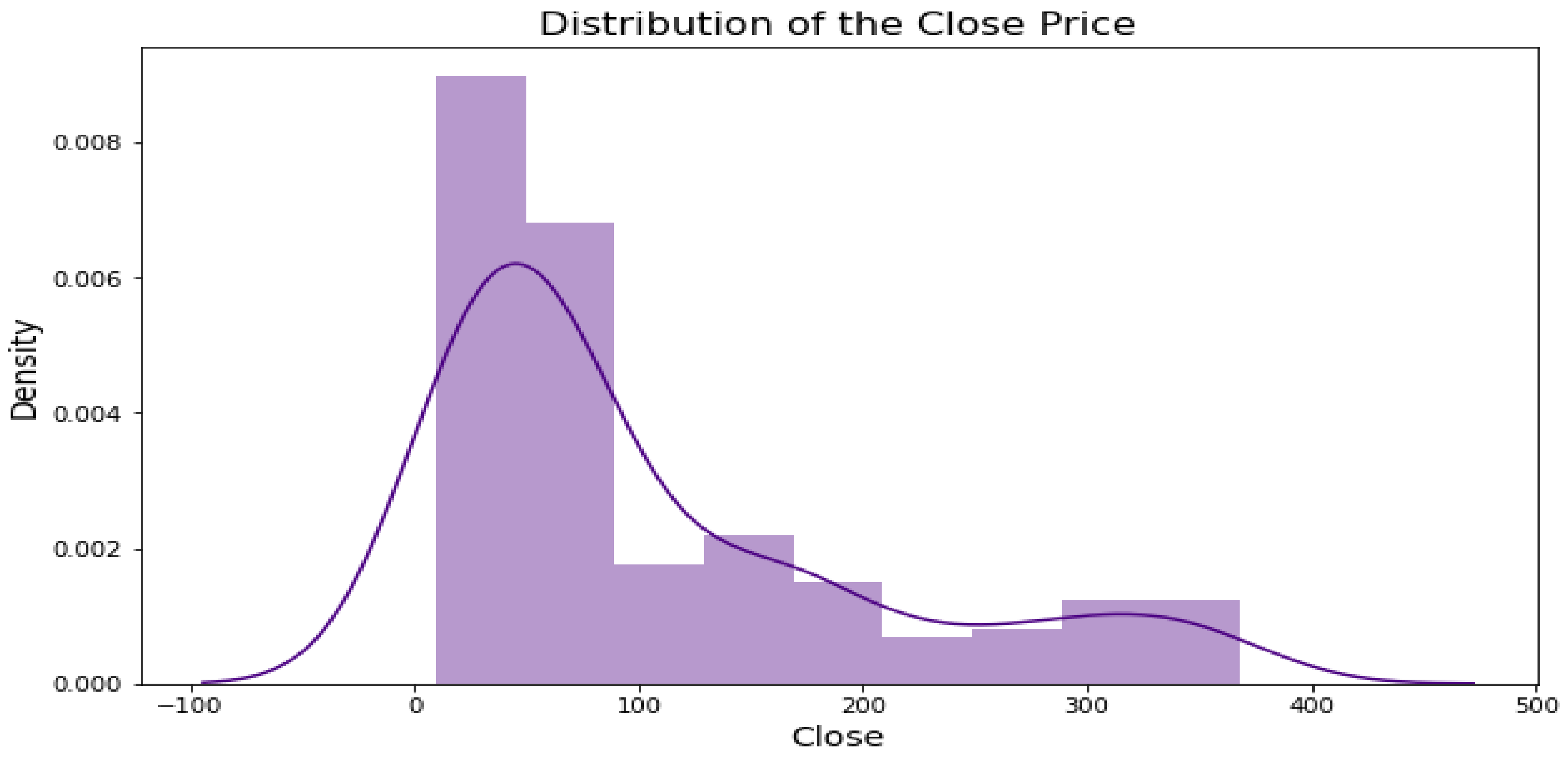


# Distribution of ‘Open’



# Distribution of 'Close' (Dependent)

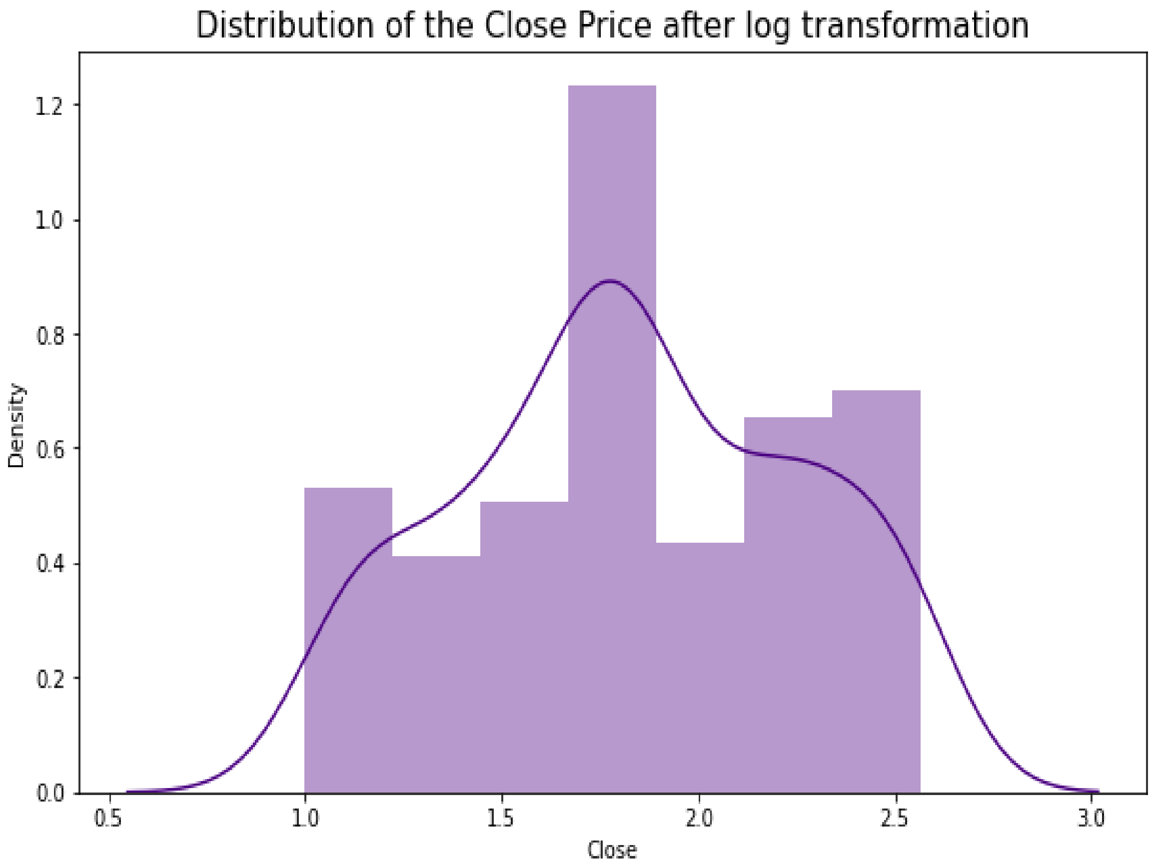


# Transformation of Data

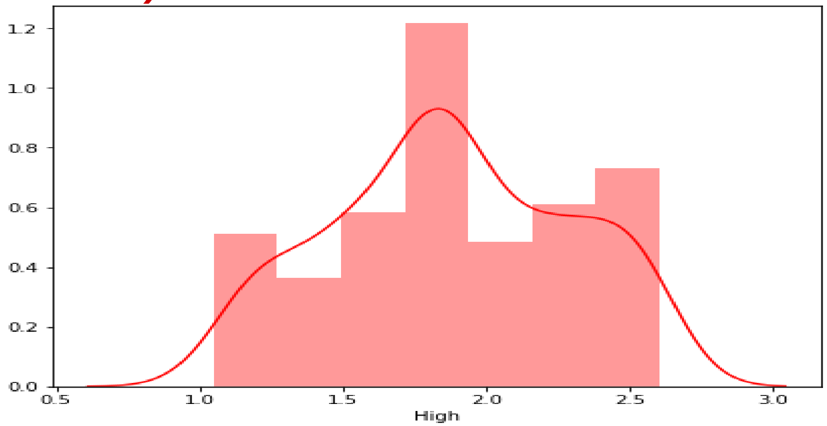
## Data Transformation

As observed in the preceding slides, the observed data was found to be skewed. We will transform the data to make it uniform before passing it into our machine learning models. Let's have a look at how they will look once the transformation is applied to them.

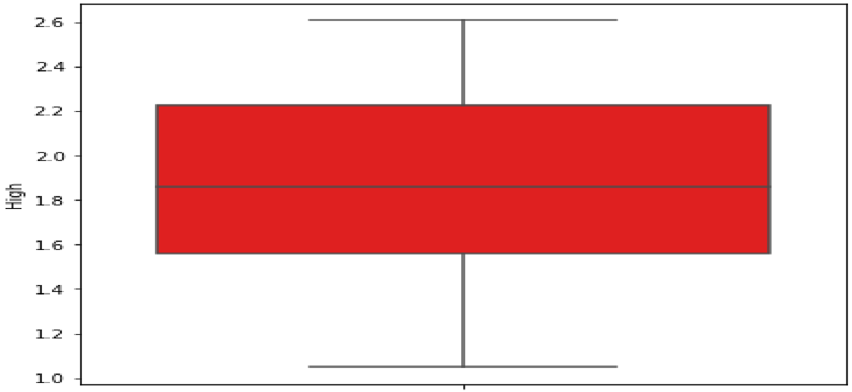
The image on the right shows how the distribution of our close price would look after a log transformation is applied to it.



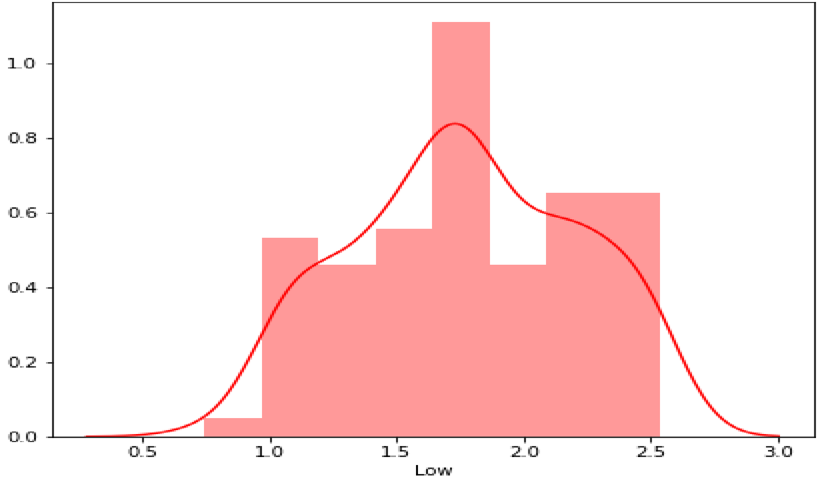
# (Cont..)



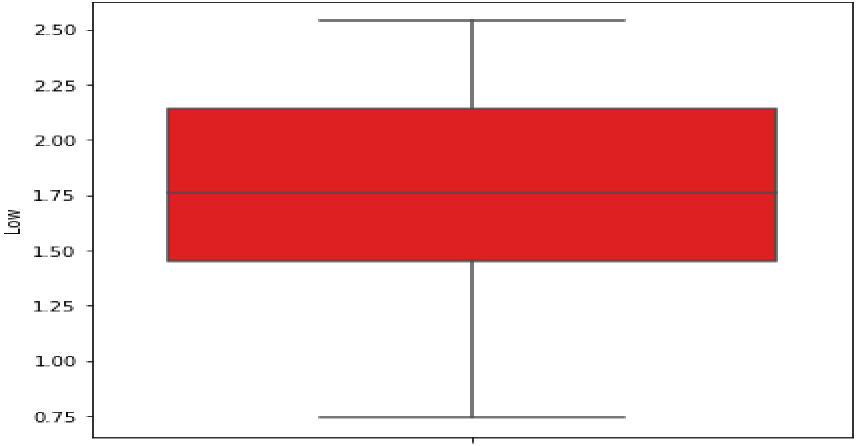
0.0
0.2
0.4
0.6
0.8
1.0
1.2
0.5
1.0
1.5
2.0
2.5
3.0
High

High
1.0
1.2
1.4
1.6
1.8
2.0
2.2
2.4
2.6

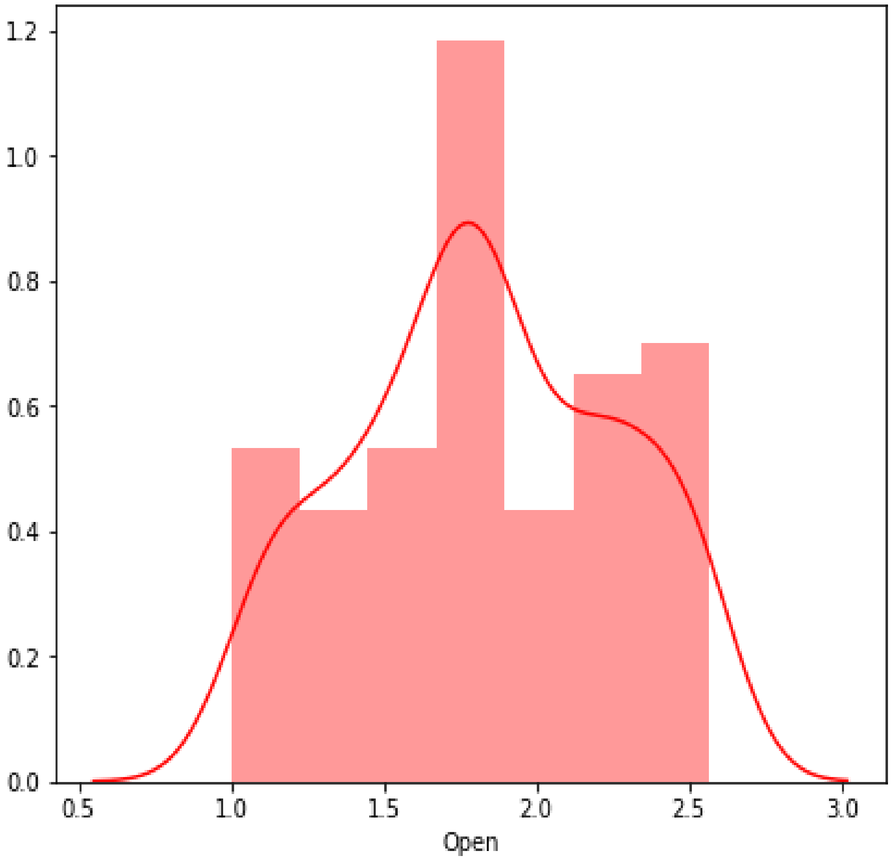



0.0
0.2
0.4
0.6
0.8
1.0
0.5
1.0
1.5
2.0
2.5
3.0
Low

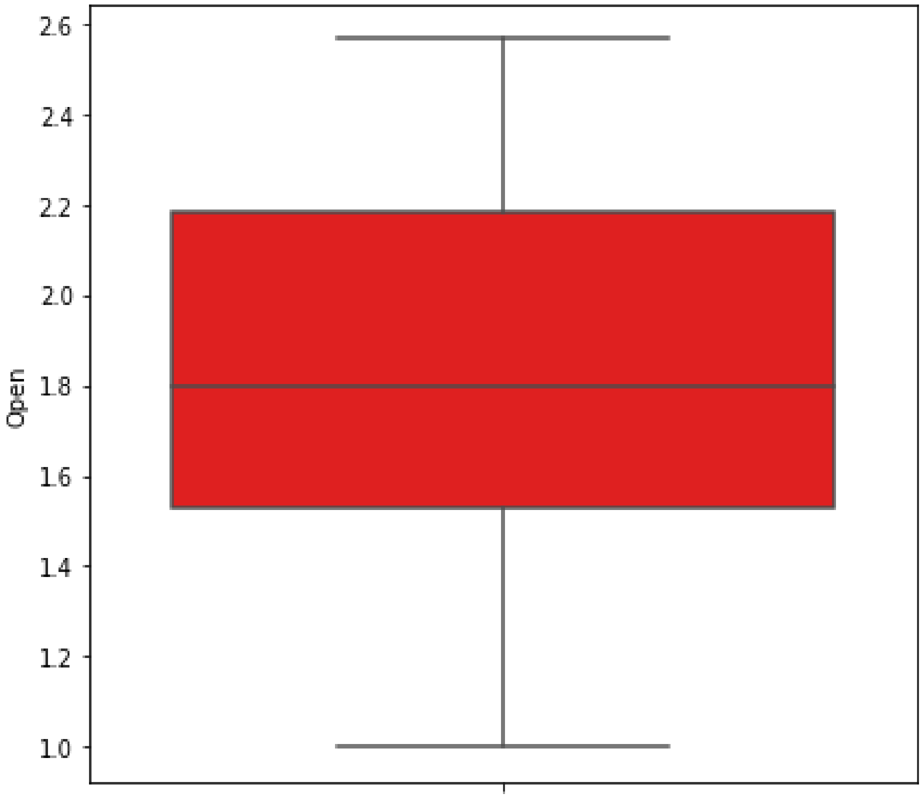



Low
0.75
1.00
1.25
1.50
1.75
2.00
2.25
2.50

# (Cont..)

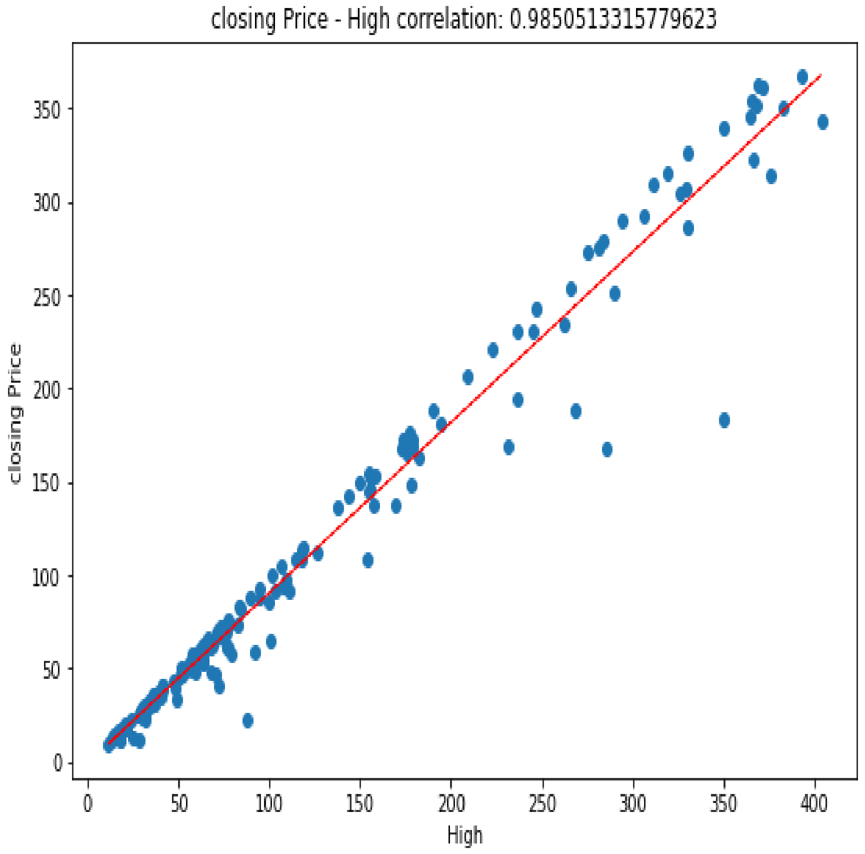


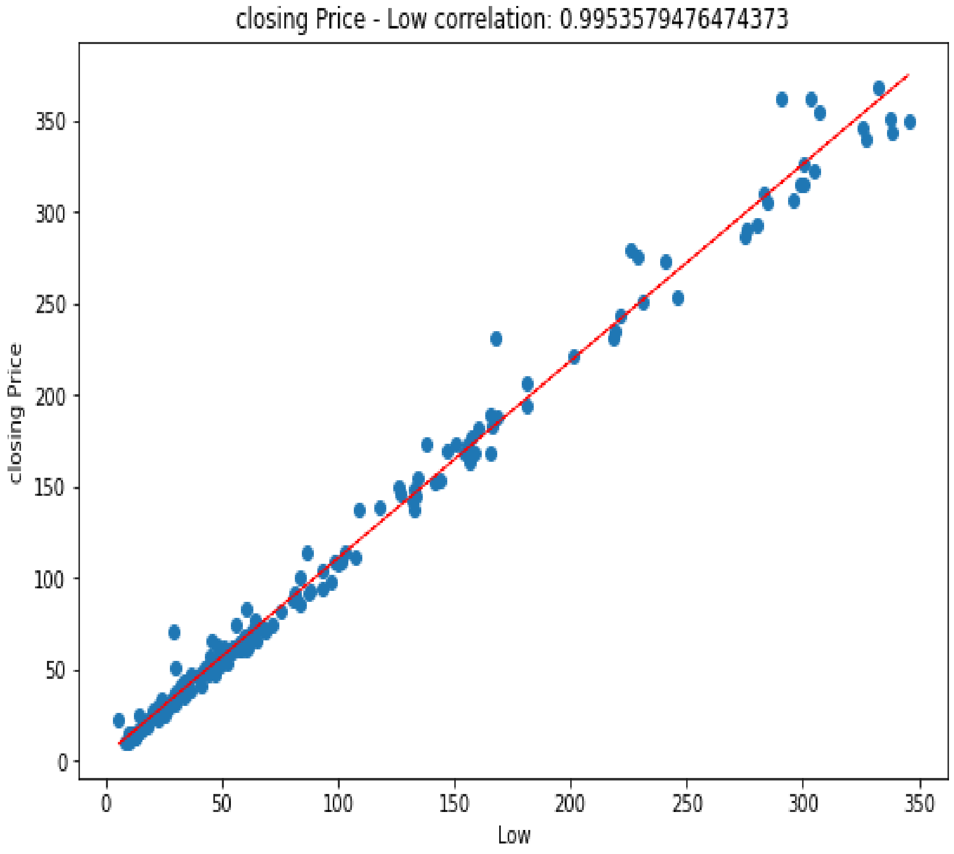
1.2
1.0
0.8
0.6
0.4
0.2
0.0
0.5
1.0
1.5
2.0
2.5
3.0
Open

2.6
2.4
2.2
2.0
1.8
1.6
1.4
1.2
1.0
Open

# Correlation of ‘Closing Price’ with Independent Features:



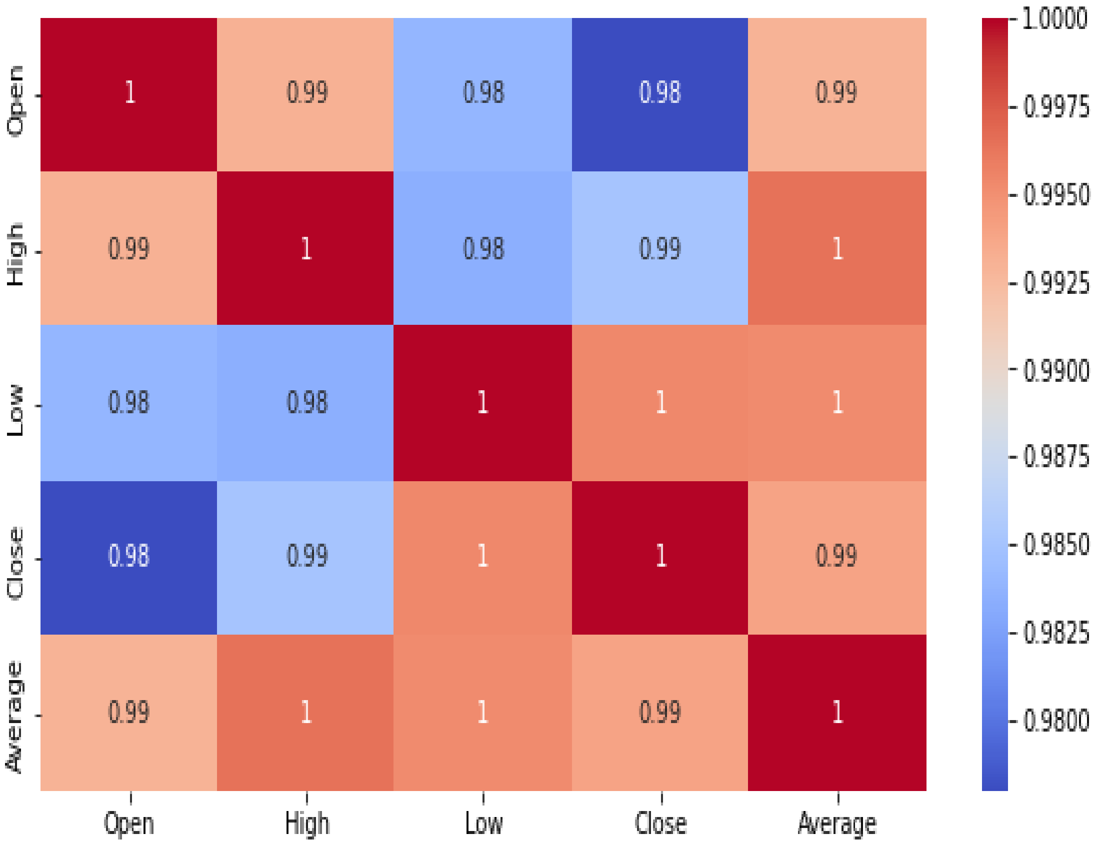
(Cont..)

As we can see that there is linear relation and high correlation between each independent variables and dependent variable.

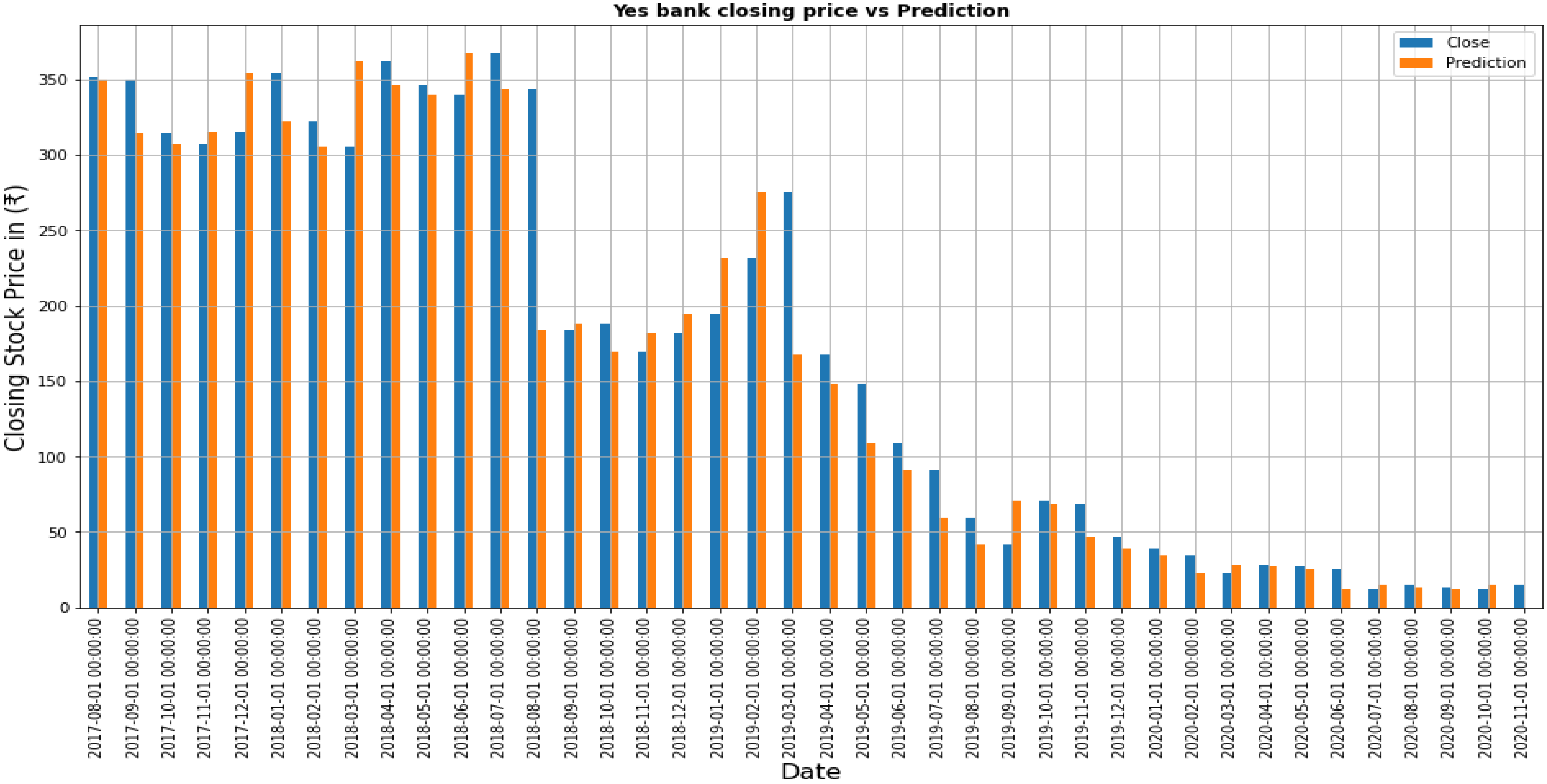

The correlation is 0.985, 0.995, 0.978 This suggests a high level of correlation, e.g. a value above 0.5 and close to 1.0.

# Correlation Matrix

- The correlation matrix helps us visualize the correlation of each parameter with respect to every other parameter.
- The shades changes from the highest to lowest (or vice versa) correlations.
- We can see in the matrix on this
slide that our dependent variable (close price) is highly correlated with all the other independent variables



# Bar Graph Comparison between Actual and predicted Price (price predicted  by Shifht() function



# Model Selection

● We passed the data into:

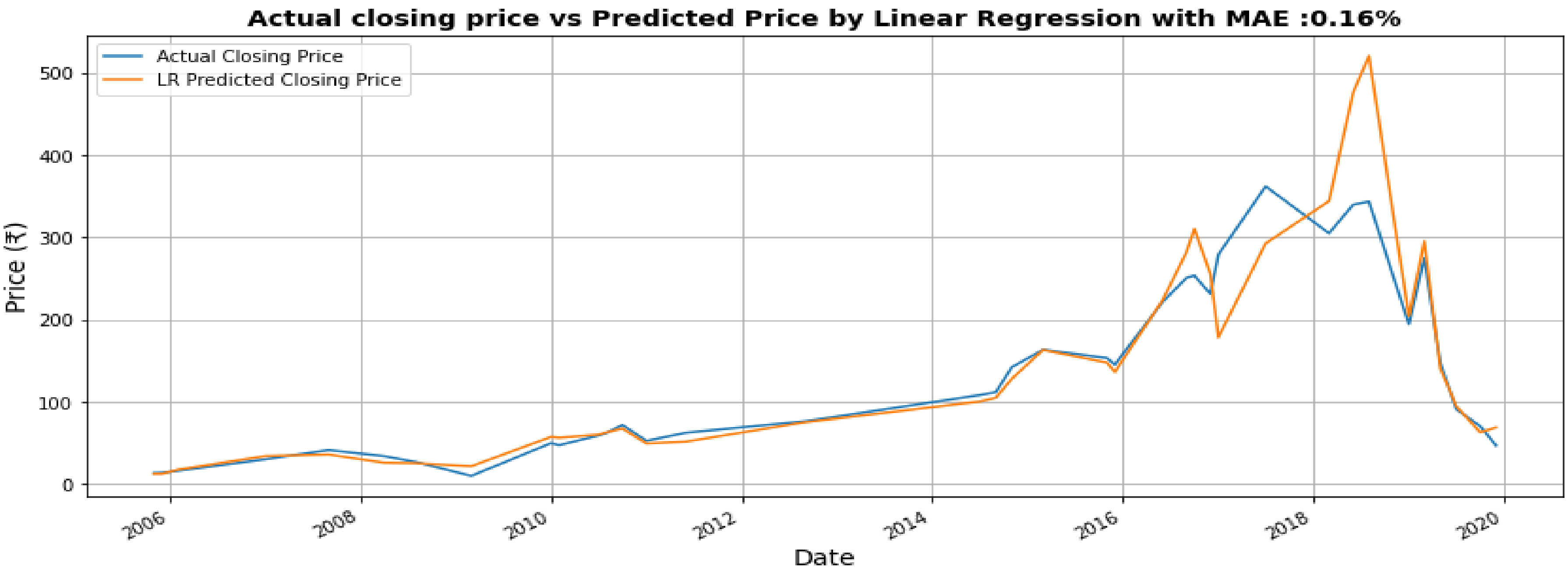
- Linear Regression Model
- Lasso Regression
- Ridge Regression
- ElasticNet

● We checked the performance of the model across various parameters.

●Then we decided our best models on the basis of following metrics

- R2
- Adjusted R2
- Training Accuracy
- Mean squared error and Root mean squared error.

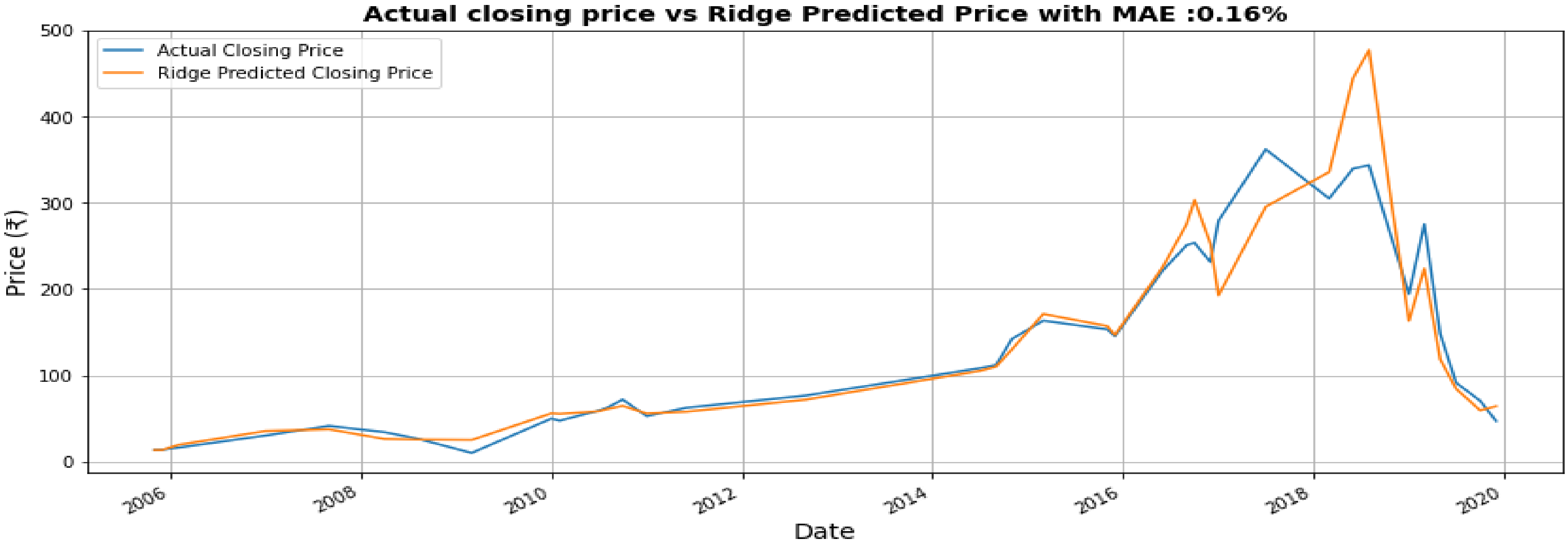
# Linear Regression




**Explanations**:

- Our Linear Model predicted the close price with 0.16% Mean Absolute error. Having training accuracy 94.03%.
- R2 tells us that our independent is able to describe 95% of our dependent variable.
- Adjusted R2 is about 91.44%, just because we consider 17 independent features adjusted R2 would be the best matrix to consider.

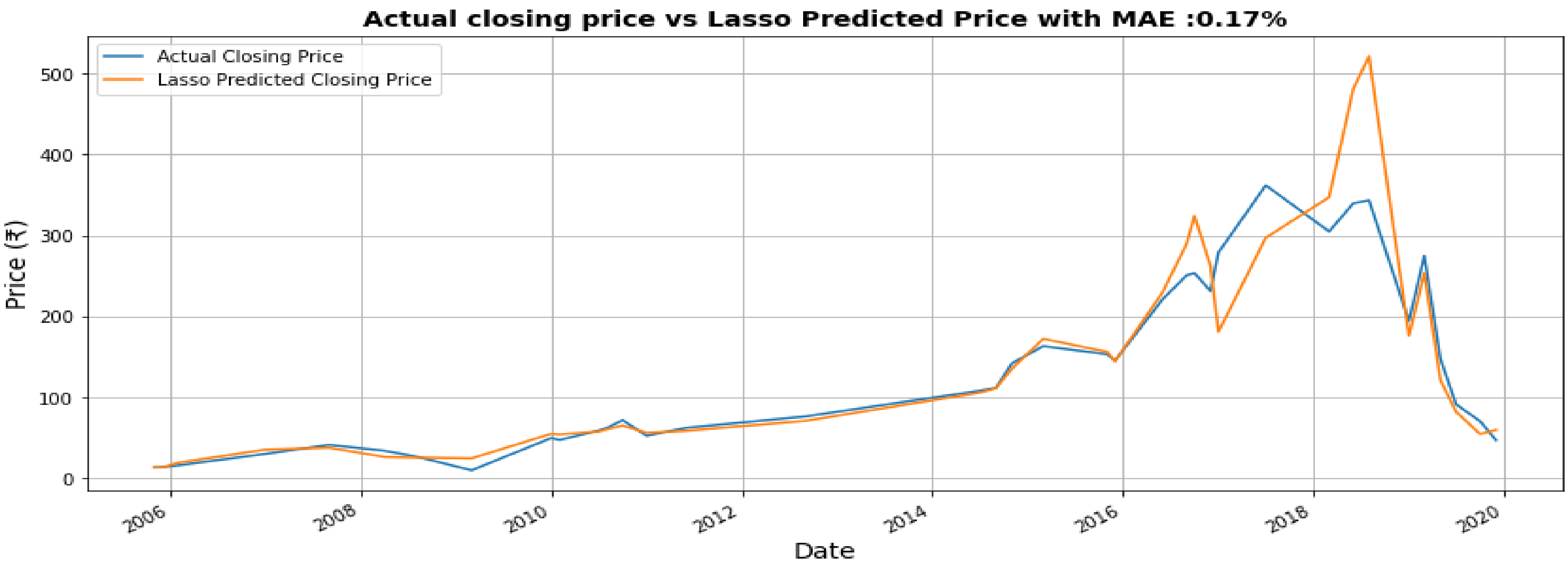
# Ridge Regression




**Explanations**:

- Our Ridge predicted the close price with 0.16% Mean Absolute error. Having training accuracy 94.57%.
- Here, R2 is about 95.25% which means model's independent features is able to describe 95.25% of our dependent variable.
- Adjusted R2 is about 91%. We'll consider adjusted R2 because we have too many independent features

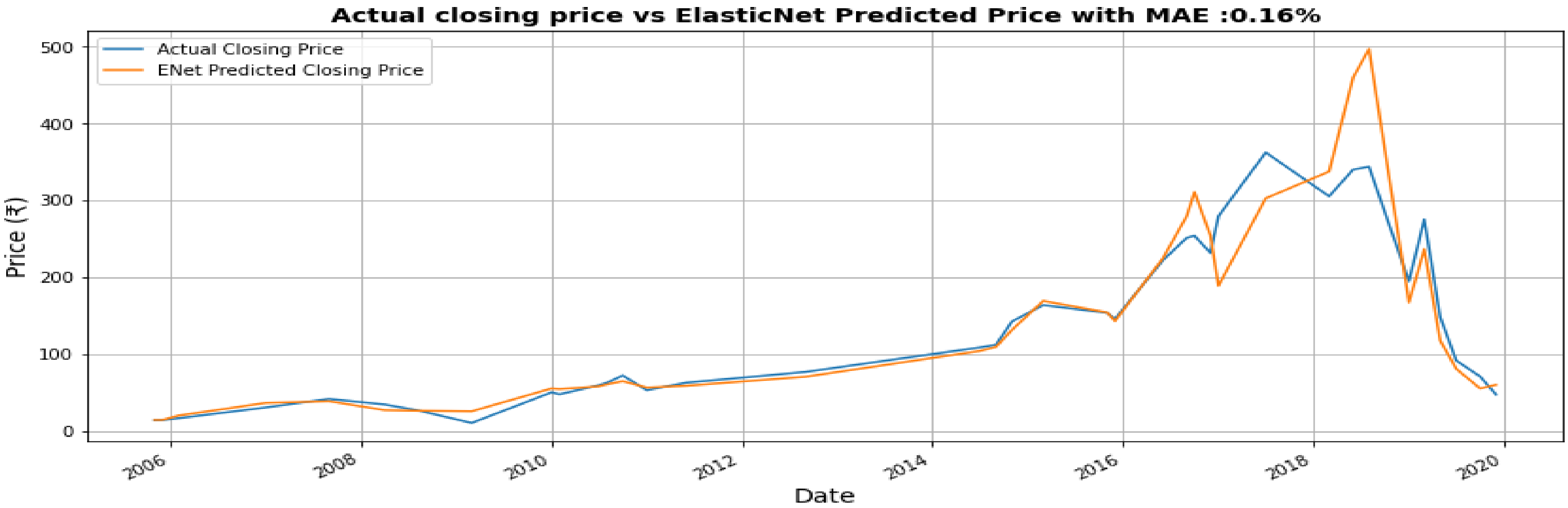
# Lasso Regression




**Explanations**:

- Lasso predicted the close price with 0.17% Mean Absolute error. Having training accuracy 94.57%.
- Here, R2 is about 94.96% which means models' independent features is able to describe 94.96% of our dependent variable.
- Adjusted R2 is about 90.46%. We'll consider adjusted R2 because we have too many independent features.

# Elastic Net Regression Before Cross Validation:

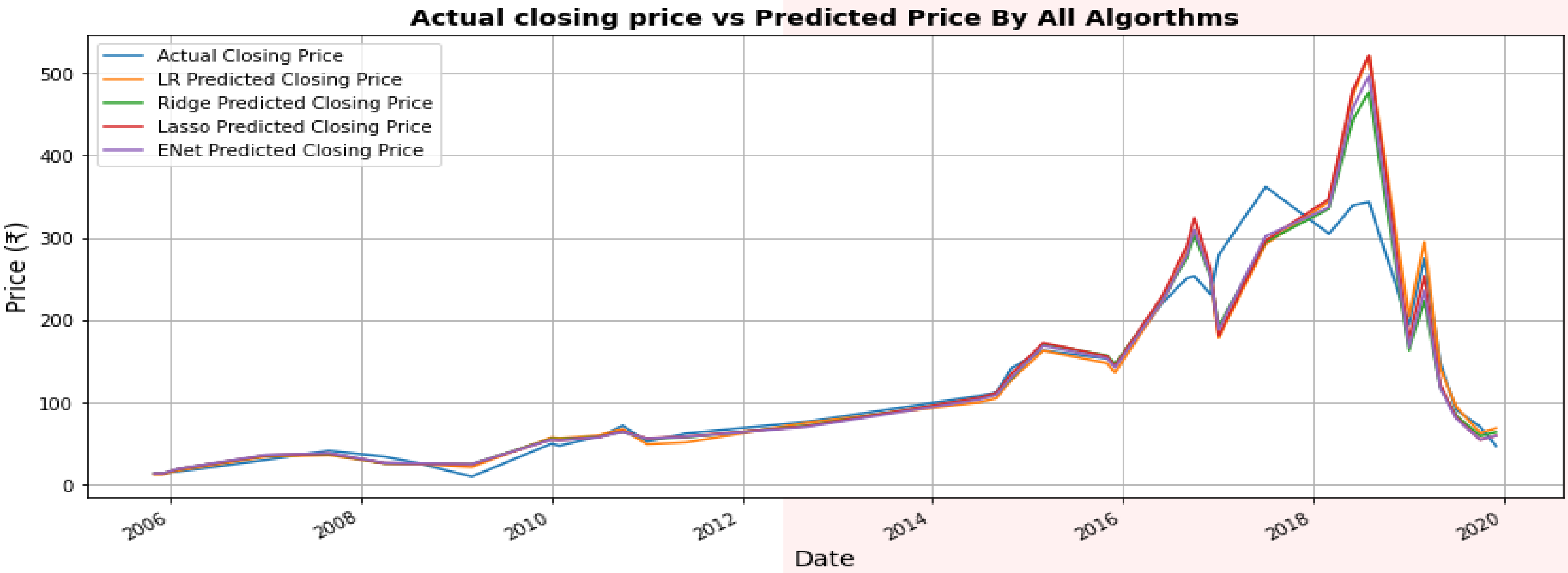



**Explanations**:

- Our Linear Model predicted the close price with 0.16% Mean Absolute error. Having training accuracy 82.64%.
- R2 tells us that our independent is able to describe 95.11% of our dependent variable.

- Adjusted R2 is about 90.74%, just because we consider 17 independent features adjusted R2 would be the best matrix to consider.

# Comparispn among all Model predictions in one graph:




**Final Explanation**: In this combined comparison graph among actual closing price and Predicted closing price predicted by all four models, Linear Regression and Lasso are predicting closing price of next month better than Ridge and ElasticNet.

All four models have good R2 and Adjusted R2.

# Final Matrix:

| | Linear Regression | Ridge | Lasso | Elastic Net |
|---|---|---|---|---|
| **MSE** | 0.008368 | 0.008848 | 0.009377 | 0.009096 |
| **RMSE** | 0.091477 | 0.094061 | 0.096833 | 0.095375 |
| **R2** | 0.955079 | 0.952505 | 0.949664 | 0.951169 |
| **Adjusted_R2** | 0.914887 | 0.910009 | 0.904627 | 0.907478 |
| **Training Accuracy** | 0.940359 | 0.740745 | 0.945777 | 0.826402 |

# Conclusion

- Target Variable is strongly dependent on Independent Variables.
- By Introducing Dummy variables makes our data free of overfitting problem and improve our training accuracy from 82% to 94%
- Linear Regression and Lasso are performing better than other models with training accuracy **94.0359%** and **94.7881%** respectively.
- Apart from Linear Regression and Lasso, Ridge and Elastic Net is also performing better but they have less training accuracy.
- Ridge and ElasticNet is performing far much better after Applying Hyperparameter Tuning and Cross validation, it is because we have small set of datasets.
- R2 and Adjusted R2 are around 95% and 91% in each model.